

# An Analysis of State-Relevance Weights and Sampling Distributions on $L_1$ -Regularized Approximate Linear Programming Approximation Accuracy

Gavin Taylor  
Connor Geer  
David Piekut

TAYLOR@USNA.EDU

United States Naval Academy, 572M Holloway Rd., Stop 9F, Annapolis, MD 21402-5002

## Abstract

Recent interest in the use of  $L_1$  regularization in the use of value function approximation includes Petrik et al.'s introduction of  $L_1$ -Regularized Approximate Linear Programming (RALP). RALP is unique among  $L_1$ -regularized approaches in that it approximates the optimal value function using off-policy samples. Additionally, it produces policies which outperform those of previous methods, such as LSPI. RALP's value function approximation quality is affected heavily by the choice of state-relevance weights in the objective function of the linear program, and by the distribution from which samples are drawn; however, there has been no discussion of these considerations in the previous literature. In this paper, we discuss and explain the effects of choices in the state-relevance weights and sampling distribution on approximation quality, using both theoretical and experimental illustrations. The results provide insight not only onto these effects, but also provide intuition into the types of MDPs which are especially well suited for approximation with RALP.

## 1. Introduction

In recent years, the Reinforcement Learning community has paid considerable attention to creating value function approximation approaches which perform automated feature selection while approximating a value function (Kolter & Ng, 2009; Johns et al., 2010; Mahadevan & Liu, 2012; Liu et al., 2012). This approach frees researchers from hand-selecting and -tuning feature sets, while greatly increasing approximation accuracy. One of these ap-

proaches,  $L_1$ -Regularized Approximate Linear Programming (RALP) (Petrik et al., 2010; Taylor & Parr, 2012) is unique in that it results in an approximation of the optimal value function and makes use of off-policy samples.

However, important aspects of RALP have not been fully explored or explained. In particular, the objective function of the linear program offers an opportunity to create a better approximation of the value function in some regions of the state space at the expense of others. Optimal policies of many realistic reinforcement learning problems heavily traffic some parts of the state space while avoiding others, making an understanding of this flexibility useful.

Therefore, in Section 4, we derive a new error bound for the RALP approximation, which is tighter than those presented by Petrik et al. (2010) and provides new insight into the result of changing the state-relevance weights in the objective function. In addition, this bound provides an insight into the types of MDPs particularly well suited to the RALP approach. Finally, this section provides evidence that rather than weighting all states equally, as was done by Petrik et al. (2010) and Taylor and Parr (2012), states should instead be weighted in proportion to the stationary distribution under the optimal policy.

Additionally, if the sampling distribution is not uniform across the state space, the approximation can be heavily affected. In realistic domains, sampling is rarely uniform. Therefore, for RALP to be appropriate for these domains, it is important to address this lack of analysis and understand how the approximation is likely to be altered due to the problem's available sampling scheme.

Section 5 then discusses the impact on the approximation if learning is performed using samples drawn other than uniformly from the state space. We demonstrate that sampling from a distribution acts as a *de facto* alteration of the objective function. We also discuss the effects of sampling distributions in light of the bounds from Section 4.

The intuition provided by Sections 4 and 5 are then demon-

strated experimentally in Section 6. Using a simple, easily visualized domain, we demonstrate the effect of the various parameters on approximation quality.

## 2. Notation and Problem Statement

In this section, we formally define Markov decision processes and linear value function approximation. A *Markov decision process* (MDP) is a tuple  $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ , where  $\mathcal{S}$  is the measurable, possibly infinite set of states, and  $\mathcal{A}$  is the finite set of actions.  $P : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$  is the transition function, where  $P(s'|s, a)$  represents the probability of transitioning from state  $s$  to state  $s'$ , given action  $a$ . The function  $R : \mathcal{S} \mapsto \mathbb{R}$  is the reward function, and  $\gamma$ , a number between 0 and 1, is the discount factor, representing the comparative desire for reward at the current time step to the desire for reward at the next time step.

We are concerned with finding a value function  $V$  that maps each state  $s \in \mathcal{S}$  to the expected total  $\gamma$ -discounted reward for the process. Value functions can be useful in creating or analyzing a policy  $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$  such that for all  $s \in \mathcal{S}$ ,  $\sum_{a \in \mathcal{A}} \pi(s, a) = 1$ . The transition and reward functions for a given policy are denoted by  $P_\pi$  and  $R_\pi$ . We denote the Bellman operator for a given policy as  $T_\pi$ , and the max Bellman operator simply as  $T$ . That is, for some state  $s \in \mathcal{S}$ :

$$T_\pi V(s) = R(s) + \gamma \int_{\mathcal{S}} P(ds'|s, \pi(s)) V(s')$$

$$TV(s) = \max_{\pi \in \Pi} T_\pi V(s).$$

We additionally denote the Bellman operator for selecting a particular action  $a$  as

$$T_a V(s) = R(s) + \int_{\mathcal{S}} P(ds'|s, a) V(s').$$

The optimal value function  $V^*$  satisfies  $TV^*(s) = V^*(s)$  for all  $s \in \mathcal{S}$ .

For simplicity, in this paper we will assume no noise exists in the MDP; results can be easily extended to noisy domains using Taylor and Parr's (2012) approach of local smoothing.

Sets of samples, therefore, are defined as  $\Sigma \subseteq \{(s, a, r, s' | s, s' \in \mathcal{S}, a \in \mathcal{A})\}$ , where  $s'$  is the state the agent arrived at given that it started in state  $s$  and took action  $a$ , and  $r = R(s)$ . An individual sample in the set  $\Sigma$  will be denoted  $\sigma$ , and an element of a sample  $\sigma$  will be denoted with superscripts; that is, the  $s$  component of a  $\sigma = (s, a, r, s')$  sample will be denoted  $\sigma^s$ .

We focus on linear value function approximation for discounted infinite-horizon problems, in which the value function is represented as a linear combination of possibly non-linear basis functions (vectors). For each state  $s$ , we define

a vector  $\Phi(s)$  of features. The rows of the basis matrix  $\Phi$  correspond to  $\Phi(s)$ , and the approximation space is generated by the columns of the matrix. That is, the basis matrix  $\Phi$ , and the approximate value function  $\hat{V}$  are represented as:

$$\Phi = \begin{pmatrix} - & \Phi(s_1) & - \\ & \vdots & \end{pmatrix} \quad \hat{V} = \Phi w.$$

This form of linear representation allows for the calculation of an approximate value function in a lower-dimensional space, which provides significant computational benefits over using a complete basis; if the number of features is small and the environment is noisy, this framework can also guard against overfitting any noise in the samples.

If we define  $\Phi$  to be overcomplete, with potentially far more features than sampled states, then to receive the above benefits we must perform feature selection. In this process, a few features are chosen from the set, the span of which will represent the available linear approximation space. We can use  $L_1$  regularization to calculate a sparse  $w$ , in which nearly all features receive a weight of 0, thereby performing automated feature selection.

## 3. Previous Work

RALP was introduced by Petrik et al. (2010) to extend the capabilities of the linear programming approach to value function approximation (d'Epenoux, 1963; Schweitzer & Seidmann, 1985; de Farias & Van Roy, 2003). Given a set of samples  $\Sigma$ , the linear program is defined as follows:

$$\begin{aligned} \min_w \quad & \rho^T \Phi w \\ \text{s.t.} \quad & T_{\sigma^a} \Phi(\sigma^s) w \leq \Phi(\sigma^s) w \quad \forall \sigma \in \Sigma \\ & \|w\|_{-1} \leq \psi, \end{aligned} \quad (1)$$

where  $\rho$  is a distribution, which we call the state-relevance weights, in keeping with the (unregularized) Approximate Linear Programming (ALP) terminology of de Farias and Van Roy (2003).  $\|w\|_{-1}$  is the  $L_1$  norm of the vector consisting of all weights excepting the one corresponding to the constant feature.

This final constraint, which contributes  $L_1$  regularization, provides several benefits. First, regularization in general ensures the linear program is bounded, and produces a smoother value function. Second,  $L_1$  regularization in particular produces a sparse solution, producing automated feature selection from an overcomplete feature set. Finally, the sparsity results in few of the constraints being active, speeding the search for a solution by a linear program solver, particularly if constraint generation is used.

Other techniques have used  $L_1$  regularization in similar ways. LARS-TD (Kolter & Ng, 2009) and LC-MPI (Johns et al., 2010) both approximate the fixed point of the  $L_1$ -regularized LSTD problem. Mahadevan and Liu (2012)

introduced the use of mirror descent, which has a computation complexity which allows it to be better suited than many other approaches for online reinforcement learning problems. These above approaches are most reliable when samples are collected on-policy. Liu et al. (2012) introduced RO-TD, which converges to an approximation of the value function of a given policy, even when trained on off-policy samples.

In contrast to the above approaches, RALP provides an approximation to the value function of the *optimal* policy, even when samples are drawn from non-optimal or random policies. Approximations produced by RALP have bounded error, and have performed well experimentally in comparison to other approaches. Finally, in noisy domains, the well-known weakness of linear programming approaches to value function approximation can be mitigated or eliminated using local smoothing (Taylor & Parr, 2012).

This previous work on RALP has largely ignored the state-relevance weights  $\rho$  in the objective function, setting  $\rho = \mathbf{1}$  without discussion. However, a change in the objective function would obviously affect the solution of the linear program. In certain practical situations this would be useful to understand. For example, consider the task of calculating a value function for an aircraft in flight. The space of possible flight attitudes, velocities, etc. is very large. However, the percentage of this space trafficked in non-catastrophic flight is small; it is likely worthwhile to improve the approximation quality in this relevant portion, at the expense of accuracy in the remainder of the space.

Some previous results exist regarding the effects of changing the state-relevance weights in the closely-related ALP (de Farias & Van Roy, 2003). However, the assumptions and types of appropriate problems are very different between ALP and RALP, making these previous results insufficient. First, ALP assumes a sample is drawn from every state-action pair, an assumption which is not required for RALP. This means it was not necessary with ALP to consider the behavior of the approximation between samples or in a continuous space. Furthermore, it was not necessary to consider the effects of sampling distributions at all. This assumption was later weakened by a followup paper (de Farias & Van Roy, 2004), but not to a degree necessary for large or continuous problems, particularly when large numbers of samples are not available. The second difference is ALP is unregularized, simplifying the definition of the feasible space of the linear program. Despite these differences, these previous results will serve as a useful guide.

Additionally, previous work does not cover the effects of sampling schemes on RALP. If states are sampled heavily in one portion of the state space, the linear program will

choose a solution which tightens constraints in that portion of the space over others. In realistic settings, it can be difficult to sample uniformly, making it especially important to understand the effect of other sampling distributions on the resulting approximate value function.

The remainder of this document fills in these gaps in the previous work.

## 4. State-Relevance Weights

The theoretical results presented on RALP in the literature thus far offer no insights into the behavior of the approximation as the state-relevance weights are altered. Therefore, it is necessary to derive new bounds for RALP which contain  $\rho$  to understand its effects.

The approach we take follows the example of a proof introduced by de Farias and Van Roy (2003) to bound the ALP approximation, but we extend it match the weaker assumptions of RALP, along with the requirement that the weights be  $L_1$  regularized.

We begin by defining the relevant notation. In the following definitions, we will use  $\Re_+$  to refer to the set of non-negative real numbers.

**Definition 1** We introduce an operator  $H$ , defined by

$$(HL)(s) = \max_{a \in \mathcal{A}} \int_{\mathcal{S}} p(s'|s, a) L(s') ds',$$

for all  $L : \mathcal{S} \rightarrow \Re_+$ .

Therefore,  $(HL)(s)$  represents the expected value of  $L$  of the *next* state if actions are chosen to maximize  $L$ .

**Definition 2** A non-negative function  $L : \mathcal{S} \rightarrow \Re_+$  is a Lyapunov function if there exists a subset of states  $\mathcal{B}$  and a  $\beta_L < 1$  such that for all  $s \in \mathcal{S} \setminus \mathcal{B}$ ,  $\gamma(HL)(s) \leq \beta_L L(s)$ .

For an example of a Lyapunov function defined over a MDP, consider the simple case of a random walk along the non-negative number line, so that  $s \in \mathbb{Z}_+$ . Assume a single action, in which the probability

$$p = p(s_{t+1} = \max(m - 1, 0) | s_t = m) > 0.5$$

and

$$p(s_{t+1} = m + 1 | s_t = m) = 1 - p.$$

If  $L(s) = s$  and  $\mathcal{B} = \{0\}$ , then  $L$  is a valid Lyapunov function, because  $L(s)$  is expected to decrease for all  $s \neq 0$ .

Lyapunov functions are often used to prove stability of Markov processes. Definition 2 differs from the definition commonly used for stability analysis in a few ways. First, in stability analysis, it is required that  $\mathcal{S}$  be countable, and

that  $\mathcal{B}$  be finite. We have made neither of these assumptions, though our bounds will be tightest when  $\mathcal{B}$  is small. The second difference is we have added a multiplicative term of  $\gamma$ . Because of these differences, a Lyapunov function as defined in Definition 2 may not strictly evidence stability.

Besides stability analysis, Lyapunov functions have also previously appeared in Reinforcement Learning literature, though in different contexts from our application (Perkins & Barto, 2003; Rohanimanesh et al., 2004).

In the remainder of this section, we will occasionally refer to the weighted max-norm, where for a vector  $U$  and a function  $F$ ,  $\|U\|_{\infty, F} = \max_i |U(s_i) \cdot F(s_i)|$ , and the weighted  $L_1$  norm, where  $\|U\|_{1, F} = \sum_i |U(s_i) \cdot F(s_i)|$ .

We will start with the following Lemma. To conserve space, and because the proof is similar to one presented by de Farias and Van Roy (2003), we reserve the proof for Appendix A.

**Lemma 1** Assume samples have been drawn from every possible state-action pair. Let  $\mathcal{W} = \{w : \|w\|_{-1} \leq \psi\}$ , and let  $w^* = \min_{w \in \mathcal{W}} \|V^* - \Phi w\|_{\infty}$ . Additionally, for a given Lyapunov function  $\Phi w_L$ , let

$$\bar{w} = w^* + \|V^* - \Phi w^*\|_{\infty, \frac{1}{\Phi w_L}} \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) w_L.$$

If a Lyapunov function  $\Phi w_L$  is constructed such that  $\bar{w} \in \mathcal{W}$ , then,

$$\|V^* - \Phi \bar{w}\|_{1, \rho} \leq \frac{2\rho^T \Phi w_L}{1 - \beta_{\Phi w_L}} \min_{w \in \mathcal{W}} \|V^* - \Phi w\|_{\infty, \frac{1}{\Phi w_L}}.$$

We note that proving the existence of a Lyapunov function as required in the above lemma is trivial. First we construct a weight vector  $w_L$  with all zeros but for a positive weight corresponding to the bias feature; this results in  $\Phi w_L$  being a valid Lyapunov function. Second, we note that in this case  $\|\bar{w}\|_{-1} = \|w^*\|_{-1}$ , meeting the requirement that  $\bar{w} \in \mathcal{W}$ .

We must now remove the assumption that a sample exists for every state-action pair. To enable us to bound the behavior of the value function between samples, we make the following assumption, similar to the sufficient sampling assumption made by Petrik et al. (2010):

**Assumption 1** Assume sufficient sampling, that is, for all  $s \in \mathcal{S}$  and  $a \in \mathcal{A}$ , there exists a  $\sigma \in \Sigma$  such that  $\sigma^a = a$  and:

$$\begin{aligned} \|\phi(\sigma^s) - \phi(s)\|_{\infty} &\leq \delta_{\phi} \\ \|R(\sigma^s) - R(s)\|_{\infty} &\leq \delta_R \\ \|p(s'|s, a) - p(s'|s, a)\|_{\infty} &\leq \delta_P \quad \forall s' \in \mathcal{S} \end{aligned}$$

This assumption is not unrealistic. For example, if the reward function, basis functions, and transition functions are Lipschitz continuous, then appropriate values of  $\delta_{\phi}$ ,  $\delta_R$ , and  $\delta_P$  are easily calculated given the greatest distance between any point in the state space and a sampled point.

We describe the maximum difference between the RALP solution and the true solution by using the limits from Assumption 1 to demonstrate the following Lemma:

**Lemma 2** Let  $M_1$  be an MDP with optimal value function  $V_1^*$ , and let  $\Sigma$  be an incomplete set of samples drawn from  $M_1$  such that not all state-action pairs are sampled, but Assumption 1 is fulfilled. Therefore, the RALP for  $M_1$  has the constraint  $T_{\sigma^a} \Phi(\sigma^s) w \leq \Phi(\sigma^s) w$  for all  $\sigma \in \Sigma$ , and the bounded  $L_1$  constraint, but is missing all other possible RALP constraints.

There exists an MDP  $M_2$  with an optimal value function  $V_2^*$ , identical in every way to  $M_1$  but for the reward function, such that the RALP solution with no missing constraints is equal to the RALP solution constructed on  $\Sigma$ , and  $\|V_1^* - V_2^*\|_{\infty} \leq \frac{2(\delta_{\phi}\psi + \delta_R + \delta_P\psi)}{1 - \gamma}$ .

**Proof Sketch:** We first show that if  $R_1$  and  $R_2$  are the respective reward functions of  $M_1$  and  $M_2$ ,

$$\|R_1 - R_2\|_{\infty} \leq 2(\delta_{\phi}\psi + \delta_R + \delta_P\psi).$$

We then show that if  $\|R_1 - R_2\|_{\infty} \leq \delta$ ,

$$\|V_1^* - V_2^*\|_{\infty} \leq \frac{\delta}{1 - \gamma}.$$

We leave the details of the proof for Appendix B.

We are now prepared to present the first result of this paper.

**Theorem 1** Let  $\Phi w_L$  be a Lyapunov function as required by Lemma 1. Define  $\Sigma$ , MDPs  $M_1$  and  $M_2$  and their respective optimal value functions  $V_1^*$  and  $V_2^*$  as in Lemma 2. Define  $\epsilon_p = \delta_{\phi}\psi + \delta_R + \delta_P\psi$ . Let  $\tilde{w}$  be the RALP solution to  $M_1$ .

$$\|V_1^* - \Phi \tilde{w}\|_{1, \rho} \leq \frac{2\rho^T \Phi w_L}{1 - \beta_{\Phi w_L}} \min_{w \in \mathcal{W}} \|V_2^* - \Phi w\|_{\infty, \frac{1}{\Phi w_L}} + \frac{2\epsilon_p}{1 - \gamma}.$$

**Proof:** Because  $\tilde{w}$  is an optimal solution given all samples from  $M_2$ , Lemmas 1 and 2 allow us

$$\|V_2^* - \Phi \tilde{w}\|_{1, \rho} \leq \frac{2\rho^T \Phi w_L}{1 - \beta_{\Phi w_L}} \min_{w \in \mathcal{W}} \|V_2^* - \Phi w\|_{\infty, \frac{1}{\Phi w_L}}.$$

Additionally, Lemma 2 gave us  $\|V_1^* - V_2^*\|_{\infty} \leq \frac{2\epsilon_p}{1 - \gamma}$ .



Because  $\rho$  is a probability distribution,

$$\|V_1^* - V_2^*\|_{1,\rho} \leq \|V_1^* - V_2^*\|_\infty \leq \frac{2\epsilon_p}{1-\gamma}.$$

Due to the triangle inequality, Theorem 1 follows.

#### 4.1. Discussion

This bound is not only tighter than those presented in previous literature, but also allows us to analyze the RALP approximation quality in new ways. First, we can observe which Lyapunov functions would result in a better approximation, and discuss the characteristics of MDPs which allow for those Lyapunov functions, and therefore lend themselves particularly well to value function approximation by RALP. Second, as  $\rho$  now appears in our bound, the bound provides a way of relating our choice of  $\rho$  to approximation quality, allowing for more intuitive and successful parameter assignments. We address these in turn.

The Lyapunov function  $\Phi w_L$  appears in the first term of our bound in three places, namely the dot product with  $\rho$ , the definition of  $\beta_{\Phi w_L}$ , and in the norm defining the “optimal”  $w$  to which we compare our approximation. We first note that the bound becomes smaller as  $\beta_{\Phi w_L}$  decreases. This suggests that the more stable the MDP, the better RALP can approximate the value function.

This interpretation of Theorem 1 leads to other intuitive explanations. Consider an MDP with only  $L(s) = 1 \forall s \in \mathcal{S}$  as a Lyapunov function. Now assume two nearby samples, one where a “good” action is taken, in the direction of positive reward, and another where a “bad” action is taken, in the direction of negative reward. When the linear program is solved, due to the proximity of the two samples, the constraint corresponding to the “bad” sample is nearly certain to be loose, and may as well be removed. However, if the MDP is highly stable, then these two extremely different samples would be unlikely, and both constraints are candidates for being tight. Therefore, more samples from an MDP with a small  $\beta_L$  are likely to be involved in defining the feasible space, potentially resulting in an improved approximation.

The appearance of the Lyapunov function in the norm of the bound indicates the bound is tighter when the feature space  $\Phi$  allows for a close approximation in areas where the Lyapunov function is small. Bertsimas et al. (1998) demonstrated that a small Lyapunov function value correlates in expectation with a higher probability in the stationary distribution of a Markov chain. This is particularly interesting when considering the appearance of the dot product between the Lyapunov function and  $\rho$ . This dot product makes it apparent that the approximation improves when  $\rho$  is large only where  $\Phi w_L$  is small. This provides evidence that the stationary distribution of the MDP under the opti-

mal policy may be an advantageous setting for  $\rho$ . This evidence meshes well with the intuition that greater accuracy is most useful in frequently-visited states.

## 5. Sampling Distribution

Imagine an MDP with a small finite state space and a single action. Ideal sampling would provide a single sample from each state, giving us an objective function of  $\sum_{s \in \mathcal{S}} \rho(s) \Phi(s)w$ . However, if sampling from a distribution across the state space, this ideal situation would be unlikely; some states would go unsampled, while others would be sampled multiple times. Because the objective function is defined on samples, this means states that were sampled multiple times would appear in the objective function multiple times, causing the linear program to tighten constraints at those states at the expense of accuracy in other states.

Of course, a similar scenario occurs in infinite state spaces as well. Multiple states near to each other may be sampled, while other regions have very few samples; this encourages the linear program to choose features and an approximate value function which tightens constraints in heavily-sampled regions at the expense of sparsely-sampled regions. In this section we discuss the effects of sampling from an arbitrary distribution over the state space  $\mu$  and ways this can help the researcher understand how to design sampling methods.

**Observation 1** *Let  $\Sigma_1$  and  $\Sigma_\mu$  be sample sets of equal cardinality  $N$  drawn from the state space from the uniform distribution and an arbitrary distribution  $\mu$ , respectively. Let  $\Phi_1$  and  $\Phi_\mu$  be the feature matrices defined over the states of sets  $\Sigma_1$  and  $\Sigma_\mu$ . For any weight vector  $w$ ,*

$$\mathbb{E} [\mu \Phi_1 w] = \mathbb{E} [\mathbf{1} \Phi_\mu w].$$

This observation is easy to support; both expectations equal  $\int_{\mathcal{S}} \mu(s) \Phi(s)w \, ds$ .

This means that for any  $w$ , sampling from a non-uniform distribution  $\mu$  provides an equivalent objective function in expectation to sampling from a uniform distribution but setting the state-relevance weights equal to  $\mu$ . We note this is not equivalent to expecting the same approximate value function as the constraints remain different; this makes the effect of altering the sampling distribution greater than that of altering the objective function alone.

Additionally, the bound presented in Theorem 1 offers an interpretation of results from sampling from a distribution. As sampling becomes less uniform,  $\epsilon_p$  likely increases, due to the existence of larger unsampled regions. For example, this would happen in the Lipschitz-continuous case for full-filling Assumption 1. However, for a sampling distribution

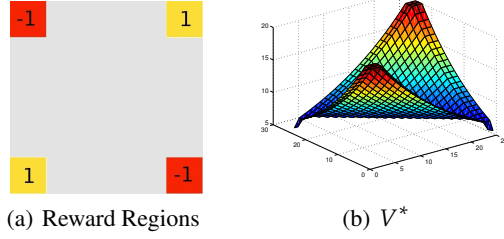


Figure 1. Room Domain

which is dense where the Lyapunov value is small, then the total Lyapunov value in the numerator of the first addend is small, as well. Therefore, it may be that the ideal sampling distribution is one which is dense where the Lyapunov value is low, but still provides sufficient coverage for  $\epsilon_p$  to be small.

## 6. Experimental Results

In this section, we demonstrate experimentally the conclusions drawn in the previous sections. Previous literature has already demonstrated RALP’s effectiveness in common benchmark domains; the purpose of this section therefore, is to clearly illustrate the conclusions of the previous sections. In a simple, easily visualized domain, we make a series of comparisons. First, we compare the approximation accuracy of sampling from a domain with a stable Lyapunov function to the accuracy resulting from sampling from a domain without such a function. Next, we compare the accuracy of the approximation resulting from sampling uniformly to the accuracy of the approximation resulting from sampling from two different nonuniform distributions. Finally, we compare the approximation accuracy of calculating an approximation with  $\rho = 1$  to the approximation accuracy of calculating an approximation when  $\rho$  is nonuniform. This is demonstrated using two different, nonuniform distributions.

The results were obtained by drawing samples and calculating an approximation 500 times for each compared approach; the error  $|V^* - \Phi w|$  was then calculated, and averaged across all 500 trials. Finally, we calculate and display the difference between the average errors from the two approaches. So, if  $\hat{V}_i^A$  is the approximation from the  $i$ -th run on approach A, then when comparing two approaches, A and B, a point on the graphs of Figure 2 equals

$$\sum_{i=1}^{500} \frac{|V^*(s) - \hat{V}_i^A(s)|}{500} - \sum_{i=1}^{500} \frac{|V^*(s) - \hat{V}_i^B(s)|}{500}.$$

### 6.1. Domain

All of the experiments were run on a domain defined by a 25 by 25 grid world. This world included four reward regions in the corners of the grid, each of which consisted of 9 states, as can be seen in Figure 1(a). Two reward regions (colored gold) had a reward of 1, while the others had a reward of -1 (colored red). The remaining states had a reward of 0. Actions were to move one square in any of the four directions, unless constrained by a wall, in which case that action would result in no movement. The discount factor  $\gamma$  was set to 0.95. For all trials, the feature set consisted of symmetric Gaussian features centered around each  $\sigma^s$  with variances of 2, 5, 10, 15, 25, 50, and 75, plus the bias feature, resulting in  $9n + 1$  features for  $n$  samples. The optimal value function can be seen in Figure 1(b).

This value function is an easy one for RALP to approximate with proper settings of the regularization parameter and sufficient sampling. The number of samples and choices of the regularization parameter  $\psi$  were therefore chosen to illustrate the differences between the results of the compared methods, not to optimize performance.

### 6.2. Lyapunov Stable Domain

First, we demonstrate the improvement in RALP’s approximation when the domain has a stable Lyapunov function. A stable Lyapunov function can be created by forcing the actor into a defined area in the state space. In order to keep the representational difficulty of the optimal value functions the same, we created a Lyapunov function by eliminating actions which move the actor further from the nearest positive reward. This preserves the optimal policy of the unaltered domain, keeping the optimal value functions identical, making approximation accuracy a fair comparison.

In the domain without a stable Lyapunov function, the actor was free to move in the state space based on a random choice among the four actions. However, in the domain with a stable Lyapunov function, the actor was only allowed to move in the two directions which would not move it further from the nearest goal. We note that not all remaining actions are optimal, so sampling still includes off-policy samples.

This creates a Lyapunov function where  $L(s)$  equals the Manhattan distance from  $s$  to the nearest of state (1,1) or (25,25), and  $\mathcal{B} = \{(1,1), (25,25)\}$ . In each trial, we uniformly sampled 20 samples. The regularization parameter  $\psi$  was set to 0.2 and  $\rho$  was 1.

The result of subtracting the average errors of the approximation from the domain with a Lyapunov function from the domain without a Lyapunov function can be seen in Figure 2(a). Therefore, positive values indicate higher error from

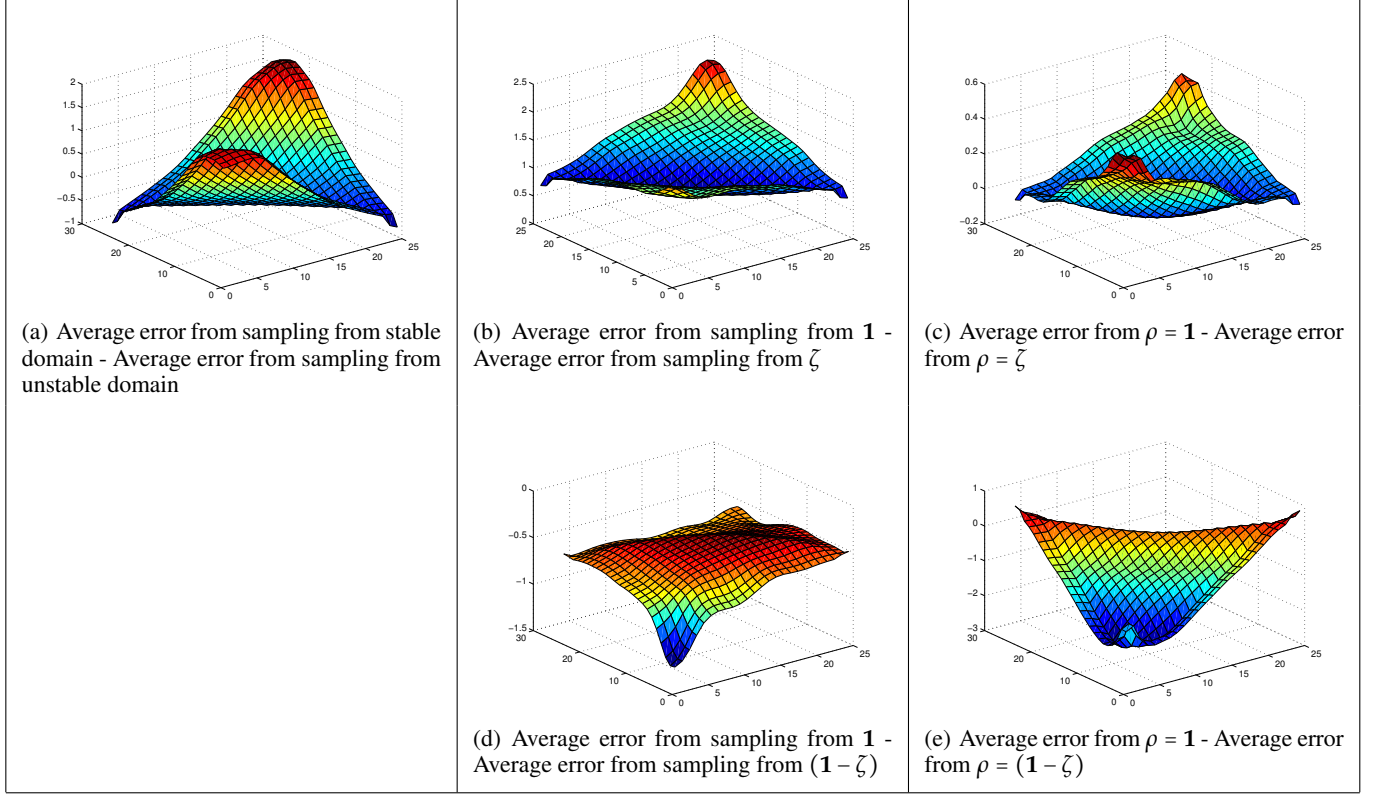


Figure 2. Difference in Average Error

the domain without a Lyapunov function.

The results show that an approximation learned from samples drawn from a stable domain is more accurate than an approximation learned from a less stable domain everywhere except for where  $s$  is near equal distance from the two goal states. This reinforces the intuition from Subsection 4.1 that samples drawn from a stable domain are more effective than those that are not, particularly near the most heavily-visited regions.

### 6.3. Sampling from a Nonuniform Distribution

Next we illustrate the change in approximation accuracy when sampling from a nonuniform distribution  $\mu$ . Section 5 presents evidence that a distribution which is most dense where  $L(s)$  is smallest may be advantageous. To create such a distribution, an agent was started at a random state, and was allowed to take the optimal policy for 25 steps. This was done 10,000 times, and the number of visits to each state was tabulated and normalized. This defined our distribution, which was heaviest on the edges and reward corners, and otherwise slightly increasing with increasing proximity to the positive reward regions. We will refer to this distribution as  $\zeta$ .

For these trials, 20 samples were drawn per run from the

domain with the stable Lyapunov function as discussed in Subsection 6.2. The regularization parameter  $\psi$  was set to 1.5, and  $\rho = 1$ . The average error from the 500 uniformly sampled runs was subtracted from the average error from the 500 runs with  $\mu = \zeta$ ; the result can be seen in Figure 2(b). Because the error from the nonuniform sampling was subtracted from the error from uniform sampling, the positive difference indicates the results from sampling from  $\zeta$  were superior. The results show the distribution met the goal from Section 5; sampling was varied enough to keep  $\epsilon_p$  low, while dense enough in the areas where  $L(s)$  was small.

We then subtracted  $\zeta$  from  $1$  and normalized, making a distribution we will refer to as  $1 - \zeta$ , which was largest where an agent was least likely to traverse in the stable domain. We subtracted the average error from sampling from  $\mu = 1 - \zeta$  from the average error from sampling uniformly, to produce Figure 2(d). A positive value would indicate larger error from the approximations on uniformly sampled states. However, through the entirety of the state space, and particularly in the most trafficked areas, sampling from  $1 - \zeta$  gave us an inferior result.

This provides evidence for the conclusions of Section 5 that a sampling distribution which is densest in the areas where the Lyapunov function is smallest would produce the best

approximations.

#### 6.4. Changing the State-Relevance Weights

Lastly, we illustrate the effect on the approximation of changing the state-relevance weights. 200 samples were drawn uniformly from the state space; whereas the effects from the previous two experiments are most pronounced when samples are sparse, the effects from altering  $\rho$  are most pronounced when a number of constraints can be tightened in a given region.

We would prefer to set  $\rho$  to the stationary distribution; however, because the domain is not recurrent, this is not an option. However, the distribution  $\zeta$  created for sampling in Subsection 6.3 is large where the Lyapunov function is small, making it a reasonable replacement. In one set of trials,  $\rho(s)$  was set to the value of this distribution at that state; in the other,  $\rho = 1$ .  $\psi$  was set to 4. Average error from the approximation resulting from a nonuniform  $\rho$  was subtracted from average error from the approximation resulting from a uniform  $\rho$ . Therefore, a positive value indicates a better approximation from the nonuniform  $\rho$ . Sampling was done uniformly from the stable domain.

Figure 2(c) shows that nearly the entire state space was more accurate with a nonuniform  $\rho$ , except for where  $\zeta(s) \approx 0$ . The difference is small because with a large number of samples, both approximations were quite accurate.

In addition, we compared the uniform  $\rho$  approximation to an approximation using  $\rho = 1 - \zeta$ , which is large where the Lyapunov value is large, resulting in an increased dot product  $\rho\Phi w$  in Theorem 1. Again, we subtracted the error of the approximation resulting from  $\rho = \zeta - 1$  from the error of the approximation resulting from  $\rho = 1$ , producing Figure 2(e). A positive value indicates the nonuniform  $\rho$  approximated that state better than did using a uniform  $\rho$ . However, there are few positive values as the use of  $\rho = 1 - \zeta$  resulted in a dramatically inferior approximation, particularly in areas where  $\rho$  was small.

From both figures, it is clear that a higher  $\rho$  value in a given portion of the state space resulted in an improved approximation, particularly if designed with Theorem 1 in mind.

## 7. Conclusion

The experimental success of RALP in previous literature, along with its easily-fulfilled assumptions, suggests promise for its application to real-life, complicated, and complex domains. Despite this promise, and despite the evidence that the effects are dramatic, no theory had been produced to analyze changes in the approximation quality given changes to the objective function parameter  $\rho$ , or due to differences in sampling strategies. These considerations

are essential to the use of RALP in the real world; it is rarely possible to sample uniformly, and the importance of accuracy across the state space is rarely consistent.

In this paper, we demonstrate the importance of understanding these ideas, and produce a bound on the approximation error of RALP which is tighter and more informative than previous bounds. This bound provides intuition into the quality of the RALP approximation as a function of state-relevance weights and sampling distributions. In addition, we demonstrated that the quality of a RALP approximation is particularly good when the domain is stable and has a Lyapunov function with a small  $\beta_L$ .

Future work remains, particularly in the area of solving the linear program quickly in the presence of large amount of data. Though convex optimization solvers are considered “fast,” with large amounts of data, the memory and time requirements may be too large for realistic use. Fortunately, it may be that the structure of the problem, the small percentage of tight constraints, and small percentage of active features will avail itself to faster, but equivalent, approaches.

## Acknowledgements

Thank you to the anonymous reviewers for their help in improving this paper. Additionally, we are grateful for support from the Naval Research Laboratory Information Management & Decision Architecture Branch (Code 5580), as well as financial support by the Office of Naval Research, grant numbers N001613WX20992 and N0001414WX20507.

## References

- Bertsimas, Dimitris, Gamarnik, David, and Tsitsiklis, John N. Geometric bounds for stationary distributions of infinite markov chains via lyapunov functions. Technical report, Massachusetts Institute of Technology, 1998.
- de Farias, Daniela Pucci and Van Roy, Benjamin. The Linear Programming Approach to Approximate Dynamic Programming. *Operations Research*, 2003.
- de Farias, Daniela Pucci and Van Roy, Benjamin. On Constraint Sampling for the Linear Programming Approach to Approximate Dynamic Programming. *Mathematics of Operations Research*, pp. 462–478, 2004.
- d’Epenoux, F. A probabilistic production and inventory problem. *Management Science*, 1963.
- Johns, Jeffrey, Painter-Wakefield, Christopher, and Parr, Ronald. Linear complementarity for regularized policy evaluation and improvement. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R.S., and Culotta, A.



(eds.), *Advances in Neural Information Processing Systems* 23, pp. 1009–1017, 2010.

Kolter, J. Zico and Ng, Andrew. Regularization and Feature Selection in Least-Squares Temporal Difference Learning. In Bottou, Léon and Littman, Michael (eds.), *Proceedings of the 26th International Conference on Machine Learning*, pp. 521–528, Montreal, Canada, June 2009. Omnipress.

Liu, Bo, Mahadevan, Sridhar, and Liu, Ji. Regularized Off-Policy TD-Learning. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2012.

Mahadevan, Sridhar and Liu, Bo. Sparse Q-Learning With Mirror Descent. In *Conference on Uncertainty in Artificial Intelligence*, 2012.

Perkins, Theodore J. and Barto, Andrew G. Lyapunov Design for Safe Reinforcement Learning. *The Journal of Machine Learning Research*, 3:803–832, 2003.

Petrik, Marek, Taylor, Gavin, Parr, Ronald, and Zilberstein, Shlomo. Feature selection using regularization in approximate linear programs for markov decision processes. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Rohanimanesh, Khashayar, Platt, Robert, Mahadevan, Sridhar, and Grupen, Roderic. Coarticulation in Markov Decision Processes. In *Advances in Neural Information Processing Systems*, 2004.

Schweitzer, Paul J. and Seidmann, Abraham. Generalized Polynomial Approximations in Markovian Decision Processes. *Journal of mathematical analysis and applications*, 110(6):568–582, 1985.

Taylor, Gavin and Parr, Ronald. Value function approximation in noisy environments using locally smoothed regularized approximate linear programs. In de Freitas, Nando and Murphy, Kevin (eds.), *Conference on Uncertainty in Artificial Intelligence*, pp. 835–842, Catalina island, California, June 2012.

## A. Proof of Lemma 1

As stated in Section 4, this proof is very similar to Theorem 3 by de Farias and van Roy (2003), but we include it for clarity nonetheless. The Lemma constructs a point in the feasible space of the linear program which provides an approximation with bounded error, and then shows the point chosen by RALP must be no further than that constructed point. The proof first requires a series of additional Lemmas.

**Lemma 3** For any functions  $V$  and  $\bar{V}$ ,

$$|T\bar{V} - TV| \leq \gamma \max_{\pi} P_{\pi} |\bar{V} - V|.$$

**Proof:** For any  $V$  and  $\bar{V}$ ,

$$\begin{aligned} T\bar{V} - TV &= \max_{\pi} (R + \gamma P_{\pi} \bar{V}) - \max_{\pi} (R + \gamma P_{\pi} V) \\ &= R + \gamma P_{\pi_{\bar{V}}} \bar{V} - R - \gamma P_{\pi_V} V \\ &\leq \gamma \max_{\pi} P_{\pi} (\bar{V} - V) \\ &\leq \gamma \max_{\pi} P_{\pi} |\bar{V} - V|, \end{aligned}$$

where  $\pi_V$  and  $\pi_{\bar{V}}$  represent the greedy policies with respect to value functions  $V$  and  $\bar{V}$ . By reversing the terms, we can show  $TV - T\bar{V} \leq \gamma \max_{\pi} P_{\pi} |\bar{V} - V|$ , leading to our result. ■

**Lemma 4** For any vector  $L$  with positive components and any vector  $V$ ,

$$TV \leq V + (\gamma HL + L) \|V - V^*\|_{\infty, \frac{1}{L}}.$$

**Proof:** Note that

$$|V^*(s) - V(s)| \leq \|V - V^*\|_{\infty, \frac{1}{L}} V(s).$$

Because of Lemma 3,

$$\begin{aligned} |(TV)(s) - (TV^*)(s)| &\leq \gamma \max_{\pi} \sum_{s' \in \mathcal{S}} P_{\pi}(s, s') |V(s') - V^*(s')| \\ &\leq \gamma \|V - V^*\|_{\infty, \frac{1}{L}} \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_a(s, s') L(s') \\ &= \gamma \|V - V^*\|_{\infty, \frac{1}{L}} (HL)(s). \end{aligned}$$

Define  $\epsilon = \|V - V^*\|_{\infty, \frac{1}{L}}$ .

$$\begin{aligned} TV(s) &\leq V^*(s) + \gamma \epsilon (HL)(s) \\ &\leq V(s) + \epsilon L(s) + \gamma \epsilon (HL)(s). \end{aligned}$$

■

**Lemma 5** Let  $w_L$  be a weight vector such that  $\Phi w_L$  is a Lyapunov function,  $w$  be an arbitrary weight vector, and

$$\bar{w} = w + \|V^* - \Phi w\|_{\infty, \frac{1}{\Phi w_L}} \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) w_L.$$

Then,  $T\Phi \bar{w} \leq \Phi \bar{w}$ .

**Proof:** Let  $\epsilon = \|V^* - \Phi w\|_{\infty, \frac{1}{\Phi w_L}}$ . For any state  $s \in \mathcal{S}$ ,

$$\begin{aligned}
 & |(T\Phi\bar{w})(s) - (T\Phi w)(s)| \\
 &= \left| T \left[ \left( \Phi w + \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) \Phi w_L \right) \right] (s) - (T\Phi w)(s) \right| \\
 &\leq \gamma \max_{\pi} \sum_{s' \in \mathcal{S}} P_{\pi}(s, s') \\
 &\quad \cdot \left| \left( \Phi w(s') + \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) \Phi w_L(s') \right) - \Phi w(s') \right| \\
 &\leq \gamma \max_{\pi} \sum_{s' \in \mathcal{S}} P_{\pi}(s, s') \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) (\Phi w_L)(s') \\
 &= \gamma \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) (H\Phi w_L)(s).
 \end{aligned}$$

The first line is a replacement of  $\bar{w}$  with its definition, the second is due to Lemma 3, the third is due to the cancellation of the two  $\Phi w$  terms and the fact that because  $\Phi w_L$  is a Lyapunov function,  $\left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) > 0$  (note this is true for states in sets  $\mathcal{B}$  and  $\mathcal{S} \setminus \mathcal{B}$  as defined in Definition 1), and the final line is due to Definition 2.

From this, we can conclude

$$T\Phi\bar{w} \leq T\Phi w + \gamma \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) H\Phi w_L.$$

We can apply Lemma 4 to get

$$T\Phi\bar{w} \leq \Phi w + \epsilon(\gamma H\Phi w_L + \Phi w_L),$$

and therefore,

$$\begin{aligned}
 T\Phi\bar{w} &\leq \Phi w + \epsilon(\gamma H\Phi w_L + \Phi w_L) + \gamma \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) H\Phi w_L \\
 &= \Phi w + \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) \Phi w_L - \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) \Phi w_L \\
 &\quad + \epsilon(\gamma H\Phi w_L + \Phi w_L) + \gamma \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) H\Phi w_L \\
 &= \Phi\bar{w} - \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) \Phi w_L + \epsilon(\gamma H\Phi w_L + \Phi w_L) \\
 &\quad + \gamma \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) H\Phi w_L \\
 &= \Phi\bar{w} + \epsilon(\gamma H\Phi w_L + \Phi w_L) - \epsilon \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) (\Phi w_L - \gamma H\Phi w_L) \\
 &\leq \Phi\bar{w} + \epsilon(\gamma H\Phi w_L + \Phi w_L) - \epsilon(\Phi w_L + \gamma H\Phi w_L) \\
 &= \Phi\bar{w}.
 \end{aligned}$$

The penultimate line can be shown given that  $\Phi w_L - \gamma H\Phi w_L > 0$  and

$$\begin{aligned}
 \frac{2}{1 - \beta_{\Phi w_L}} - 1 &= \frac{2}{1 - \max_{s \in \mathcal{S} \setminus \mathcal{B}} ((\gamma(H\Phi w_L)(s)) / ((\Phi w_L)(s)))} - 1 \\
 &= \max_{s \in \mathcal{S} \setminus \mathcal{B}} \frac{(\Phi w_L)(s) + \gamma(H\Phi w_L)(s)}{(\Phi w_L)(s) - \gamma(H\Phi w_L)(s)}
 \end{aligned}$$

Lemma 5 demonstrates that all constraints in RALP will be satisfied by  $\bar{w}$ , with the exception of the constraint enforcing the  $L_1$  regularization. However, we have required even this constraint to be satisfied by requiring  $\bar{w} \in \mathcal{W}$ . Therefore,  $\bar{w}$  lies in the feasible region for RALP.

**Lemma 6** *If every state-action pair is represented with a constraint in the RALP, a vector  $\bar{w}$  solves the RALP if and only if it solves*

$$\begin{aligned}
 &\underset{\bar{w}}{\operatorname{argmin}} \quad \|V^* - \Phi\bar{w}\|_{1,\rho} \\
 &s.t. \quad T_a\Phi(s)\bar{w} \leq \Phi(s)\bar{w} \quad \forall s \in \mathcal{S}, a \in \mathcal{A} \\
 &\quad \|\bar{w}_{-1}\|_1 \leq \psi
 \end{aligned}$$

**Proof:** For any policy  $\pi$ , the Bellman operator  $T\pi$  is a contraction in max norm. If the Bellman error is one-sided,  $T$  is also monotonic. Therefore, for any  $V$  such that  $V \geq TV$ ,

$$V \geq TV \geq T^2V \geq V^*.$$

Therefore, any  $w$  that is a feasible solution to a RALP satisfies  $\Phi w \geq V^*$ . From this, we can conclude

$$\begin{aligned}
 \|V^* - \Phi w\|_{1,\rho} &= \sum_{x \in \mathcal{S}} \rho(x) |V^*(x) - \Phi(x)w| \\
 &= \rho^T \Phi w - \rho^T V^*.
 \end{aligned}$$

Because  $V^*$  is constant, minimizing  $\rho^T \Phi w$  with RALP constraints is equivalent to minimizing  $\|V^* - \Phi w\|_{1,\rho}$  with RALP constraints. ■

Given Lemmas 5 and 6, we can finally prove Lemma 1.

$$\begin{aligned}
 \|V^* - \Phi\bar{w}\|_{1,\rho} &\leq \|V^* - \Phi\bar{w}\|_{1,\rho} \\
 &= \sum_{s \in \mathcal{S}} \rho(s) |V^* - (\Phi\bar{w})(s)| \\
 &= \sum_{s \in \mathcal{S}} \rho(s) (\Phi w_L)(s) \frac{|V^* - (\Phi\bar{w})(s)|}{(\Phi w_L)(s)} \\
 &\leq \left( \sum_{s \in \mathcal{S}} \rho(s) (\Phi w_L)(s) \right) \max_{s' \in \mathcal{S}} \frac{|V^* - (\Phi\bar{w})(s')|}{(\Phi w_L)(s')} \\
 &= \rho^T \Phi w_L \|V^* - \Phi\bar{w}\|_{\infty, 1/\Phi w_L} \\
 &\leq \rho^T \Phi w_L (\|V^* - \Phi w^*\|_{\infty, 1/\Phi w_L} + \|\Phi\bar{w} - \Phi w^*\|_{\infty, 1/\Phi w_L}) \\
 &\leq \rho^T \Phi w_L (\|V^* - \Phi w^*\|_{\infty, 1/\Phi w_L} \\
 &\quad + \|V^* - \Phi w^*\|_{\infty, 1/\Phi w_L} \left( \frac{2}{1 - \beta_{\Phi w_L}} - 1 \right) \|\Phi w_L\|_{\infty, 1/\Phi w_L}) \\
 &= \frac{2\rho^T \Phi w_L}{1 - \beta_{\Phi w_L}} \|V^* - \Phi w^*\|_{\infty, 1/\Phi w_L}.
 \end{aligned}$$

The penultimate line is due to the definition of  $\bar{w}$ , and the final line occurs because  $\|\Phi w_L\|_{\infty, 1/\Phi w_L} = 1$ . ■

## B. Proof of Lemma 2

When a constraint does not exist in RALP for some state, this does not mean the value at that state is completely unconstrained; because we bounded the rate of change of all components of the approximate and true value functions in Assumption 1, the existence of a constraint constructed on a nearby state means the existence of what we will call an *implied* constraint.

This lemma explicitly constructs these implied constraints, and quantifies the maximum distance from the true constraint which would have existed had that state been sampled. It does this by building two MDPs, identical in every way, but for the reward function.  $M_1$  has been incompletely sampled, with sample set  $\Sigma$ . Every state-action pair therefore has either an explicit or implied constraint in the corresponding RALP.  $M_2$ , however, has been completely sampled. Every state-action pair in the set  $\Sigma$  is identical in  $M_2$ , but all state-action pairs not in  $\Sigma$  have a sample producing a constraint identical to the implied constraints of  $M_1$ . Because the constraints are the same, the RALP solution is the same. We demonstrate the difference in the reward functions  $R_1$  and  $R_2$  is bounded, and thus, the difference in the optimal value functions  $V_1^*$  and  $V_2^*$  is bounded.

**Lemma 7** *Given an MDP  $M_1$  such that Assumption 1 is true and incomplete sample set  $\Sigma$ , an MDP  $M_2$  exists such that constructing the RALP with constraints for all state-action pairs results in an identical RALP solution to that of the RALP constructed from  $\Sigma$ , and  $\|R_1 - R_2\|_\infty \leq 2(\delta_\Phi\psi + \delta_R + \delta_P\psi)$ .*

**Proof:** Consider an arbitrary state-action pair  $s, a$ , which is not represented by a sample in  $\Sigma$ . This means we are missing the constraint

$$R_1(s) + \gamma \sum_{x \in \mathcal{S}} [p(x|s, a)\Phi(x)]w \leq \Phi(s)w. \quad (2)$$

Let us refer to the sample in  $\Sigma$  which fulfills the sampling assumption with  $s$  and  $a$  as  $\sigma$ . We can now construct a bound for how incorrect each component of this constraint can be if we use the constraint at  $\sigma$  and our sampling assumption to replace the missing constraint. For instance, the reward function  $R(s)$  is easily bounded.

$$R_1(\sigma^s) - \delta_R \leq R_1(s) \leq R_1(\sigma^s) + \delta_R$$

We now bound  $\Phi(s)w$ . Because the sampling assumption allows each basis function to change only a finite amount, and because  $\|w_{-1}\|_1 \leq \psi$ , and  $\mathbf{1}(s) = \mathbf{1}(\sigma^s)$ ,

$$\Phi(\sigma^s)w - \delta_\Phi\psi \leq \Phi(s)w \leq \Phi(\sigma^s)w + \delta_\Phi\psi$$

The final component is  $\gamma \sum_{x \in \mathcal{S}} p(x|s, a)\Phi(x)w$ , which expresses our expected value at the next state. It will be convenient to separate the bias feature  $\mathbf{1}$  from the rest of  $\Phi$ .

We will denote the remainder of the design matrix as  $\Phi_{-1}$ , and the weights that correspond to  $\Phi_{-1}$  as  $w_{-1}$ . Similarly, we will denote the weight corresponding to  $\mathbf{1}$  as  $w_1$ .

$$\begin{aligned} \sum_{x \in \mathcal{S}} p(x|s, a)\Phi(x)w &= \sum_{x \in \mathcal{S}} p(x|s, a)w_1 + \sum_{x \in \mathcal{S}} p(x|s, a)\Phi_{-1}(x)w_{-1} \\ &= w_1 + \sum_{x \in \mathcal{S}} p(x|s, a)\Phi_{-1}(x)w_{-1} \end{aligned}$$

Again, we have bounded the allowable change in our expression of probability.

$$\begin{aligned} w_1 + \sum_{x \in \mathcal{S}} p(x|s, a)\Phi_{-1}(x)w_{-1} \\ \leq w_1 + \sum_{x \in \mathcal{S}} [p(x|\sigma^s, \sigma^a) + \delta_P] \Phi_{-1}(x)w_{-1} \\ = w_1 + \sum_{x \in \mathcal{S}} p(x|\sigma^s, \sigma^a)\Phi_{-1}(x)w_{-1} + \delta_P \sum_{x \in \mathcal{S}} \Phi_{-1}(x)w_{-1} \end{aligned}$$

Because each basis function  $\Phi$  is can be standardized such that  $\|\Phi\|_1 = 1$ , and because  $\|w_{-1}\|_1 \leq \psi$ , the second summation can be at most  $\psi$ . So,

$$\begin{aligned} \sum_{x \in \mathcal{S}} p(x|\sigma^s, \sigma^a)\Phi(x)w - \delta_P\psi &\leq \sum_{x \in \mathcal{S}} [p(x|s, a)\Phi(x)]w \\ &\leq \sum_{x \in \mathcal{S}} p(x|\sigma^s, \sigma^a)\Phi(x)w + \delta_P\psi. \end{aligned}$$

We now combine these results, and construct our implied constraint to take the place of the missing constraint expressed by Equation 2. We see that the maximum possible change by the approximate value function is  $\delta_\Phi\psi + \delta_R + \delta_P\psi$ . So, the total cumulative error in the constraint is at most  $2(\delta_\Phi\psi + \delta_R + \delta_P\psi)$ . So, we effectively have the following constraint:

$$R_1(s) + q - \gamma \sum_{x \in \mathcal{S}} [p(x|s, a)\Phi(x)]w \geq \Phi(s)w,$$

where  $|q| \leq 2(\delta_\Phi\psi + \delta_R + \delta_P\psi)$ .

Let  $M_2$  be an MDP which is identical in every way to  $M_1$ , except  $R_2(s) = R_1(s) + q$ . The RALP solution for  $M_1$  will be equivalent to the RALP solution for  $M_2$ , and  $\|R_1 - R_2\|_\infty \leq 2(\delta_\Phi\psi + \delta_R + \delta_P\psi)$ . ■

**Lemma 8** *Let  $M_1$  and  $M_2$  be MDPs that differ only in their reward vectors  $R_1$  and  $R_2$ . Let  $V_1^*$  and  $V_2^*$  be their optimal value functions. Then, for  $\|R_1 - R_2\|_\infty \leq \delta$ ,  $\|V_1^* - V_2^*\|_\infty \leq \frac{\delta}{1-\gamma}$ .*

**Proof:** Let  $s$  be an arbitrary point in the sets  $\mathcal{S}_1$  and  $\mathcal{S}_2$ , and define  $r_{1i}(s)$  and  $r_{2i}(s)$  to be the  $i$ -th reward received in exploring  $M_1$  and  $M_2$  from state  $s$ , respectively. Note that

$$V_1^*(s) = \sum_{i=0}^{\infty} \gamma^i \mathbb{E}[r_{1i}(s)]$$

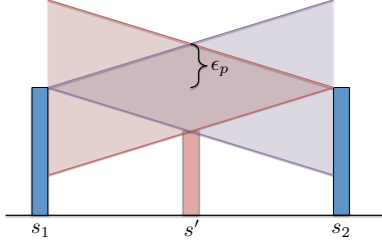


Figure 3. An illustration of Lemma 2. The blue bars are constraints at sampled points  $s_1, s_2 \in \mathcal{S}$ . The red and purple lines indicate the maximum rate of change of the value function, given our settings of  $\psi, \delta_\phi, \delta_R$ , and  $\delta_P$ . The center diamond is therefore the feasible area for the approximate value function, and the red bar is the *implied* constraint at some novel point  $s' \in \mathcal{S}$ . Because  $\epsilon_p = \delta_\phi \psi + \delta_R + \delta_P \psi$  is the maximum change, we see that the difference between the best possible setting of  $\Phi(s')w$  and the worst possible setting of  $\Phi(s')w$  is at most  $2\epsilon_p$ .

and

$$\begin{aligned} V_2^*(s) &\leq \sum_{i=0}^{\infty} \left( \gamma^i \mathbb{E} [r_{1i}(s) + \delta] \right) \\ &= \sum_{i=0}^{\infty} \gamma^i \mathbb{E} [r_{1i}(s)] + \sum_{i=0}^{\infty} \gamma^i \delta \end{aligned}$$

Therefore,

$$\begin{aligned} |V_1^*(s) - V_2^*(s)| &\leq \sum_{i=0}^{\infty} \gamma^i \mathbb{E} [r_{1i}(s)] - \left( \sum_{i=0}^{\infty} \gamma^i \mathbb{E} [r_{1i}(s)] + \sum_{i=0}^{\infty} \gamma^i \delta \right) \\ &= \sum_{i=0}^{\infty} \gamma^i \delta \\ &= \frac{\delta}{1-\gamma} \end{aligned}$$

Because this is true for an arbitrary  $s$ ,  $\|V_1^* - V_2^*\|_\infty \leq \frac{\delta}{1-\gamma}$ .

■

Lemma 2 is trivially proven by combining Lemmas 7 and 8, and is illustrated by Figure 3.